Data Wrangling (daw)
Cédric Huwyler
FS2021

Fachhochschule
Nordwestschweiz

Minichallenge: Preprocessing von Immobilienpreisen

1 Ausgangslage

Du bist Data Scientist bei einer Firma im Immobilienbereich. Im Rahmen eines Projekts im Sinne der 'Digitalisierung' sollen Hauspreis-Schätzungen durch ExpertInnen der Firma bis zu einem gewissen Masse für Vorabklärungen automatisiert werden. Dazu soll ein Machine Learning - Modell erstellt werden, das aufgrund von verschiedenen gesammelten Features den Verkaufspreis eines Hauses schätzen soll. Deine Firma plant den Kauf von entsprechenden Daten (Homegate-Daten, Location Scores, ..), möchte aber zuerst an einer einfachen Fallstudie die Chancen und Risiken abschätzen.

Deine Vorgesetzte hat dazu das *Melbourne Housing Market* - Datenset[1] entdeckt und dir ein CSV-File mit Daten übergeben, dass du untersuchen und bearbeiten sollst, um erstens entsprechendes Know-How anzusammeln und zweitens der Firma einen ersten Eindruck vermitteln zu können, welche Aufwände und unvorhergesehene Probleme speziell in der Datenaufbereitung entstehen können und wie genaue Schätzungen hier ungefähr zu erwarten sind. Du freust dich darüber, endlich wieder einmal deine Skills in Data Wrangling und explorativer Datenanalyse zu schärfen, insbesondere weil du mehrere ganze Tage Zeit dafür bekommst.

Das Datenset ist im Trainingscenter verfügbar. Versuche, dich so tief wie möglich in die einzelnen Punkte der Aufgabenstellung einzuarbeiten. Die Auseinandersetzung mit einem komplexen Datenset ist sehr aufwändig, aber essentiell, wenn du dich später erfolgreich mit einer Modellierung auseinandersetzen möchtest. Tust du dies nicht, kannst du relativ einfach an einzelnen Artefakten im Datenset scheitern.

[1] `https://www.kaggle.com/anthonypino/melbourne-housing-market`

2 Aufgabenstellung

1. Lies das Datenset ein. Verschaffe dir einen ersten Überblick über die verschiedenen enthaltenen Informationen, indem du die Dokumentation der einzelnen Spalten anschaust und den Typ der verschiedenen dort beschriebenen Merkmale bestimmst. Sind sie diskret oder stetig? Sind sie nominal-, ordinal-, intervall- oder verhältnisskaliert? Passe den Datentyp für jede Spalte entsprechend an, indem du wo nötig kategorische Variablen (`pd.Categorical` in Pandas respektive `factor` in R) einführst und Datentypen geeignet anpasst, so dass eine maximale Information über die Art der Daten bereits im Data Frame verfügbar ist und später zu keinen Unklarheiten führt. Entferne Spalten, die du für die Modellierung des Hauspreises gar nicht einsetzen willst.
2. Verschaffe dir einen Überblick über die in den verschiedenen Spalten enthaltenen Werte: Welcher Anteil der Werte fehlt? Wie sind die Werte verteilt? Was ist der Zusammenhang der Spalte mit dem Hauspreis, denkst du diese Spalte würde positiv zur Modellierung des Hauspreises beitragen? Gibt es Ausreisser oder Datenfehler, die du besser entfernst, weil sie ein später eingesetztes Modell verwirren könnten? Gibt es Wege, die verschiedenen Ausprägungen von kategorischen Variablen für das Modell geeigneter zu machen?
3. Fehlen Werte in den einzelnen Spalten zufällig oder fällt dir ein Muster auf? Macht es mehr Sinn, die Zeilen mit den fehlenden Werten zu entfernen oder die Werte geeignet zu imputieren? Gibt es Spalten mit vielen fehlenden Werten, die vielleicht ganz entfernt werden sollten? Überlege dir geeignete Strategien, um fehlende Werte zu ersetzen. Reicht die Imputation mit dem Durchschnitt oder einem konstanten Wert? Welche Vor- und Nachteile bringt eine modellbasierte Imputation? Probiere eine modellbasierte Imputation aus.
4. Gibt es starke Korrelationen zwischen einzelnen Spalten, die dein Modell verwirren könnten? Überlege dir geeignete Arten, Korrelationen zwischen zwei stetigen, zwei diskreten oder gemischten Variablen zu visualisieren oder zu quantifizieren.
5. Fasse deine Erkenntnisse in einer Preprocessing-Pipeline zusammen. Diese soll aus einer Funktion bestehen, die als Hauptargument den Pfad zum Ursprungsdatenset bekommt und über weitere Argumente (möglicherweise mit Defaultwerten) Anweisungen bekommt, wie das Datenset bereinigt werden soll: Ab welchen Werten handelt es sich bei den einzelnen Spalten um Ausreisser? Sollen die fehlenden Werte in den einzelnen Spalten entfernt oder ersetzt werden? Wie? Die Preprocessing-Pipeline soll die Daten vom CSV einlesen, korrekte Datentypen setzen (insbesondere kategorische Variablen definieren) und dann Schritt für Schritt die Daten bereinigen. Falls plötzlich ein neues, grösseres Datenset zur Verfügung steht, soll sie dieses schnell und effizient aufbereiten können.

Falls du dich bereits mit Machine Learning auseinandersetzt:

1. Macht es Sinn, einzelne Features zu skalieren, z.B. mit dem Logarithmus oder allgemein mit einer Box-Cox- oder Yeo-Johnson-Transformation? Für welche Modelltypen könnte eine solche Skalierung etwas bringen?
2. Trainiere ein lineares Regressionsmodell und ein Random-Forest-Modell auf dem bereinigten Datensatz. Welche Performance (R^2, RMSE, MAPE, ..) erwartest du ungefähr? Wird deine Erwartung erfüllt? Welches performt besser? Hast du kategorische Variablen geeignet für das Modell vorbereitet, versteht es also ihre kategorische Natur und ob sie nominal- oder ordinalskaliert sind?
3. Untersuche den Impact der verschiedenen Hyperparameter deiner Pipeline (Ausreisser entfernen, Ersetzung von fehlenden Werten, etc.) auf dein lineares Regressionsmodell und dein Random-Forest-Modell. Wie performen diese jeweils auf einem vorbereiteten Hold-Out-Set? Werden sie insbesondere von Ausreissern beeinflusst? Spielt die gewählte Strategie im Umgang mit fehlenden Werten eine Rolle für die resultierenden Performance-Masse?